A GIR Architecture with Semantic-flavored Query Reformulation

Nuno Cardoso

(preview of GIR 2010 presentation of the paper with Mário J. Silva)

GreaSE meeting - 26th January, 2010

Overview

- Motivation
- Reasoning towards the proposed GIR architecture
- GIR Architecture Overview
  - Handling queries
  - Handling documents
  - Retrieving documents
- Current challenges
- Prototype development status

Motivation

- Queries have *entities*, and entities have semantic information.
- Term-statistic query reformulation works at term level, not entity level.

Motivation (cont.)

- In order to faithfully reformulate queries regarding the user information need, we need a semantic query reformulation approach that uses such semantic information from entities and their relationships in its reformulation strategy.
- To validate this theory, let's apply it to a GIR prototype and measure its performance over geographic queries

Reasoning towards the proposed GIR architecture

- 1. To perform semantic query reformulation, we need to detect and ground entities, and have access to more information about them;

- 2. To do that, we need a knowledge base of entities and easy access to third party knowledge bases (Wikipedia, DBpedia, geographic ontologies, etc);

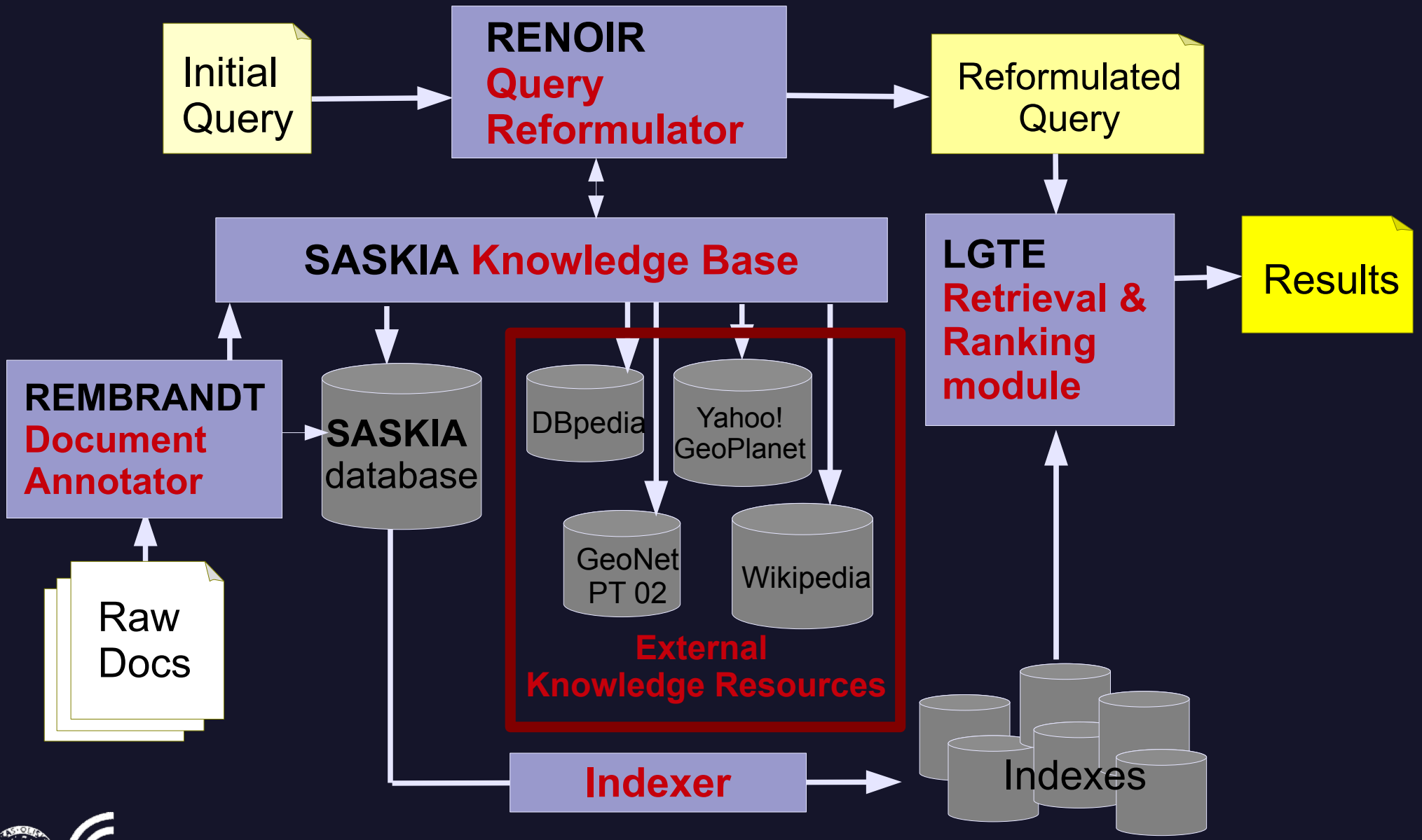
Requirements for the proposed GIR architecture (cont.)

- 3. Once query entities are grounded and the information need is captured, the retrieval phase should adapt to each query types, and use this information to rank documents (or "*one size doesn't fit all*");
- 4. We need a retrieval & ranking module that weights documents regarding its terms, entities, geoscope and temporal scope;

Requirements for the proposed GIR architecture (cont.)

- 5. Finally, we need all collection documents to have their entities, geoscope and temporal scope grounded and indexed.

GIR Architecture



Handling queries

1. RENOIR grounds all entities and detect the user intentions – geoscope, expected answer type (EAT), etc

“ *Music festivals* *in* *Germany* ”

Role: **EAT**
Ground: DBpedia:Category:Music_festivals

Role: Geoscope
Ground: DBpedia:Germany_(Country)
WOEID: 23424829

About: http://dbpedia.org/resource/Category:Music_festivals
An Entity in Data Space: dbpedia.org

Property	Value
rdf:type	skos:Concept
rdfs:label	Music festivals
skos:broader	dbpedia:Category:Live_music dbpedia:Category:Music_events dbpedia:Category:Festivals
skos:prefLabel	Music festivals
is skos:broader of	dbpedia:Category:International_music_festivals

<place yahoo:uri="http://where.yahooapis.com/v1/place/23424829" xml:lang="en">
<woeid>23424829</woeid>
<placeTypeName code="12">Country</placeTypeName>
<name>Germany</name>
<country type="Country" code="DE">Germany</country>
<admin1/>
<admin2/>
<admin3/>
<locality1/>
<locality2/>
<postal/>
<centroid>

Handling queries (cont.)

2. Use the SASKIA knowledge base to add concrete answers, use them as expanded terms for selected indexes

Initial query

term: music festivals in Germany

Final query

term: music festivals germany “wacken open air” “zappanale” “summerjan” “summer breeze open air”
event: “Wacken Open Air” “Zappanale” “Summerjan” “Summer Breeze Open Air”
place: Germany
geoscope: Germany@WOEID-23424829

Handling documents

Document tagged by REMBRANDT

Wacken Open Air takes place annually in the small town of **Wacken**, in **Germany**. **Wacken** is a metal festival.

1. REMBRANDT recognizes all named entities from documents, grounds them to entities, maps places to geoscopes, stores everything into SASKIA database.

Named Entity

Terms	Classif.
Wacken Open Air	[EVENT]
Wacken	[EVENT]
Wacken	[PLACE]
Germany	[PLACE]

SASKIA database

Entity

dbpedia:	dbpedia-owl:
Wacken_Open_Air	MusicFestival
Wacken%2C_Schleswig-Holstein	Place
Germany	Country

Geoscope

Name	WOEID
Wacken	703113
Germany	23424829

Handling documents (cont.)

2. Index generation. There are 3 kinds of indexes:

a) Term index – classic inverted index for document terms;

b) Named entity index - inverted index for named entity terms, one for each HAREM's* semantic classification;

c) Signature index - indexes of geographic and temporal signatures of documents.

*HAREM is a evaluation contest for named entity recognition systems. More information in http://www.linguateca.pt/HAREM/

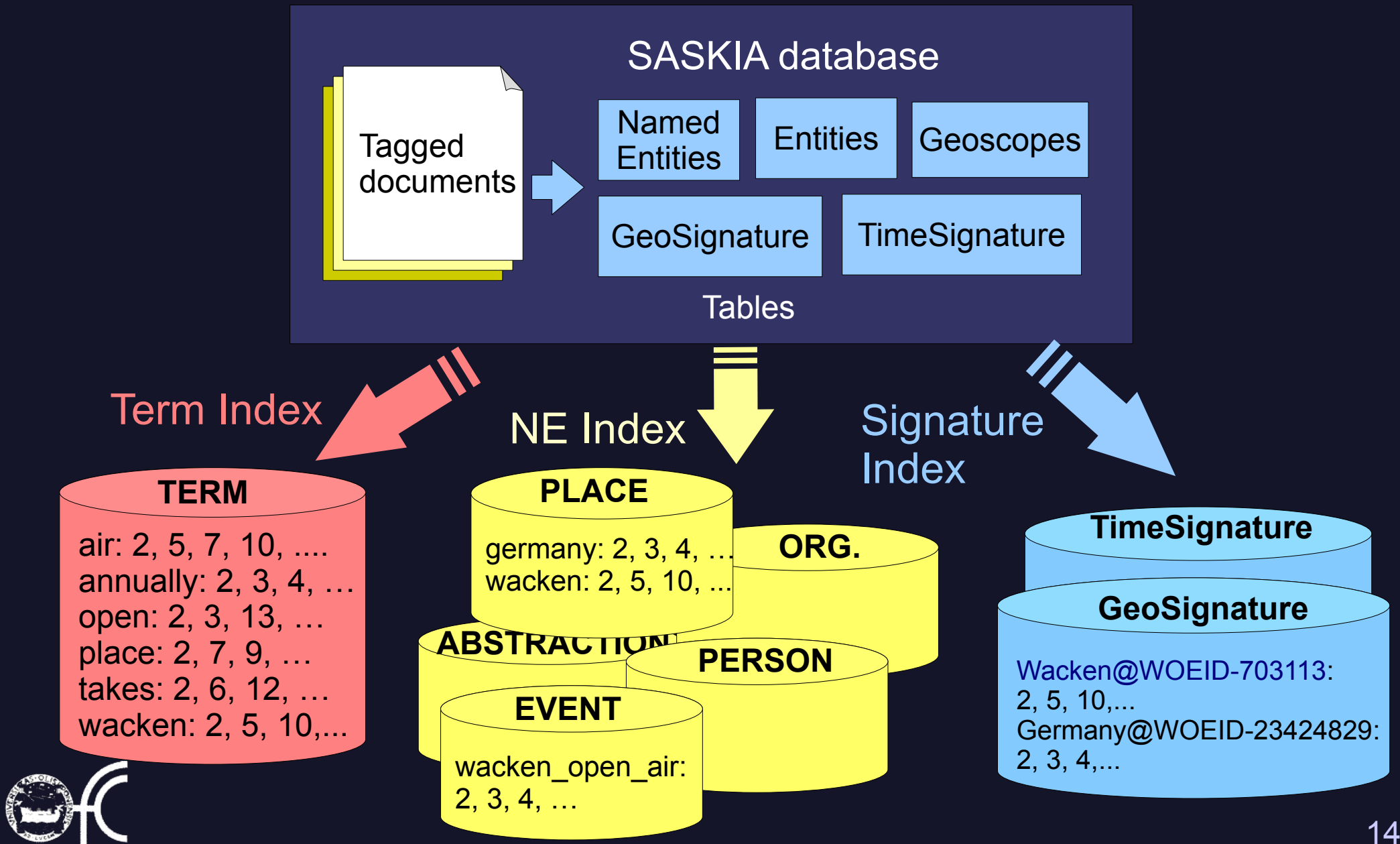
Document signatures

Surrogate of entities referred in the document that describe its scope.

```
<GeoSignature totalcount="4">
  <Doc id="571048" lang="en">
    <Place count="2" woeid="1467052">
      <NE>Harare</NE>
     <Entity>Harare</Entity>
     <Type>@HUMAN</Type>
     <Subtype>@DIVISION</Subtype>
    <DBpediaClass>Area</DBpediaClass>
    <Ancestor>Harare</Ancestor>
    <Ancestor>Harare</Ancestor>
    <Ancestor>Zimbabwe</Ancestor>
  </Place>
  <Place count="2" woeid="23425004">
     (...)
  </Place>
</GeoSignature>
```

```
<TimeSignature>
 <Doc id="523634" lang="en">
<Time count="1">
 <NE id="3645">2006</NE>
<TG>!:Y+2006</TG>
<Index>2006</Index>
</Time>
<Time count="1">
<NE id="3646">27th May, 2006</N
<TG>!:Y+2006M05S27</TG>
<Index>20060527</Index>
</Time>
(...)
</TimeSignature>
```

Handling documents (cont.)



Retrieving documents

RENOIR's reformulated query

term: music festivals germany “wacken open air” “zappanale” “summerjan” “summer breeze open air”

event: “Wacken Open Air” “Zappanale” “Summerjan” “Summer Breeze Open Air”, place: Germany

geoscope: Germany@ WOEID-23424829

Current challenges

RENOIR:

- Rule set chains for entity detection in queries for simple and complex queries
- Picking the best strategy for query reasoning:
  - What knowledge resources should we use?
  - Mapping relationships to DBpedia properties (ex: “born in” → dbpedia-owl:birthplace)
  - Handle low recall results (if “*Romanian writers born in Bucharest*” returns few or no answers, is there a plan B?)

Current challenges (cont.)

REMBRANDT / SASKIA:

- Tagging documents and populating DB is slow, complex and requires supervising
- Example: NYT collection (2002-2005)

Nr of **documents**	315.371
Nr of **named entities**	17.952.142
Nr of **classifications** assigned for**named entities**	18.364.572
Nr of classifications grounded to **entities**	3.344.235
Nr of classifications grounded do a **place**	588.621
Nr of docs with non-empty **GeoSignature**	202.624 (64%)
Nr of docs with non-empty **TimeSignature**	70.403 (22%)
Total **entities** in Saskia DB	37.001
Total **geoscopes** in Saskia DB	8.741

Current challenges (cont.)

Integration with LGTE:

- Picking the best index weights for each query (maybe learn 2 ranking?)
- Compare models (LM, BM25, DFR)

Prototype development status

- Participation in NTCIR's GeoTemporal Retrieval evaluation task (with a close collaboration with Jorge Marchado, LGTE developer)
  - 25 geographic and temporal flavored queries, as in "*Where and when did Astrid Lindgren die?*"
- Follow up: compare "footprint vs footprint-less" GIR strategies (Jorge Machado's GIR system uses geographic footprints to compute geographic ranking scores)

The end

A GIR Architecture with Semantic-flavored Query Reformulation

Nuno Cardoso and Mário J. Silva
Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE

ncardoso@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt